Scanning electron microscopy image representativeness: Morphological data on nanoparticles

Katarzyna Odziomek[1,2], Daniela Ushizima[2], Przemyslaw Oberbek[3], Tomasz Puzyn[1], Maciej Haranczyk[2]

[1]Laboratory of Environmental Chemometrics, Faculty of Chemistry, University of Gdansk, Wita Stwosza 63, 80-308 Gdansk, Poland

[2]Computational Research Division, Lawrence Berkeley National Laboratory, One Cyclotron Road, Mail Stop 50F-1650, Berkeley, CA 94720-8139, USA

[3]Materials Design Division, Faculty of Materials Science and Engineering, Warsaw University of Technology, Woloska 141, 02-524 Warsaw, Poland

Abstract

For decades the implementation of Quantitative structure-activity relationship (QSAR) methods has successfully aided scientists in predicting properties of new chemical compounds with a well-defined structure, based on their molecular descriptors. This approach could be extended to predict relevant properties of nanoparticles (NPs), ubiquitous in pharmaceuticals, cosmetics and even food products. Due to their nature, NPs exhibit different characteristics than bulk materials, which suggest conventional descriptors for QSAR techniques should be expanded and adapted. Furthermore, NP quantitation has relied upon imaging, e.g., scanning electron microscopy (SEM), generating large sets from which only the most representative records can be stored. We propose a framework for extracting morphological information contained in SEM images using computer vision algorithms, and converting them into numerical descriptors of particles. Furthermore, we supply a set of protocols for selecting optimal SEM images and determining the smallest representative image set for each of the morphological features. We demonstrate applications of our methodology by investigating tricalcium phosphate $Ca_3(PO4)_2$, a naturally mineral with a wide range of biomedical applications, such as bone implants.

Introduction

Nanoparticles (NP) are a form of chemicals or materials that exhibit different characteristics than the bulk form or single molecule.[1] Organic, inorganic or a composite[2], nanoparticles can take shapes of atom-thick layers[3], spheres[4], single- or multi-wall nanotubes[5], and various others, in which at least one external dimension is under 100nm. Ubiquitous in agriculture (e.g. herbicides[6]), cosmetics (e.g. sunscreen[7]), and medicine (e.g. drug delivery systems[8]), they present an ever-expanding field of research and development – it is estimated by 2019 the global market for nanomaterials will have reached $4.2 billion.[9] Due to their omnipresence, there is a great demand for a comprehensive characterization of the physical, chemical and biological properties of NPs, especially in relation to human and environmental safety. Similar needs are being already addressed in the field of single molecule research such as drug design,[10, 11] where e.g. absorption, distribution, metabolism, and excretion and toxicity are of special interest.[12]

One of the challenges in characterization of NPs is the difficulty to obtain precise structural data. Unlike crystalline materials or single molecules for which atomic positions can be determined via crystallography, NMR techniques or *ab initio* modeling, the NP-based materials are typically a distribution of particles (e.g. agglomerates thereof) of various shape, size and order.[1] Certainly, the experimental determination of some NPs' properties, e.g. aqueous solubility, does not require the knowledge of morphology of a NP sample. However, for classification or assessment of more complex qualities, such as structure-property relationships, toxicity or mode of action[13], this knowledge is *sine qua non*.

Moreover, the information about structure and/or morphology of a nanoparticle sample can open exciting opportunities to computationally model NP characteristics e.g. via quantitative structure-property/activity relationships (QSPR/QSAR) approaches. These techniques employ combinations of structural descriptors in prediction of physical-chemical and/or biological properties of chemical substances, based on statistical models derived from known data.[14] QSPR/QSAR modeling for nanoparticles ("Nano-QSAR") is an emerging field providing protocols to predict complex properties and asses environmental risks for nanoparticle-based technologies.[15] Without the knowledge of the exact structure of a given NP sample, researchers have explored two approaches to calculate structural descriptors:

(i) approximate the structure with a simple model (e.g. bulk structure or a surface models based on the latter), for which some descriptors can be calculated, and

(ii) (ii) recover partial structural information about a sample of NPs through available experimental imaging techniques, e.g. via image analysis of photomicrographs of a NP sample.

Following the approach (i), Kar and coworkers used hydrophobicity, surface charge and topological descriptors to estimate the uptake of magnetofluorescent NPs in human epithelia pancreatic cells.[16] Similarly, Gajewicz and associates utilized the enthalpy of formation and Mulliken's electronegativity to predict the toxicity of metal oxides to human keratinocyte cell line.[17]

The approach (ii) typically involves scanning electron microscopy (SEM) or transmission electron microscopy (TEM), which have become standard techniques to investigate micro- and nano-sized particle structures in the food industry (e.g. comparing milk treatments[18]), mechanics

(e.g. investigating combustion byproducts[19-23]), and medicine (e.g. examining wear particles of artificial joint fillers)[24-28]. There are various ways of examining information contained in SEM and TEM images. Intuitive approaches follow natural human visual inspection when investigating new objects: estimating size, shape, contour and texture. In other words, it is possible to characterize the NP samples in terms of length, surface area, circularity etc. [18-29] The identified morphology can be later correlated with properties. For example, Adachi *et al*. examined the relationship between soot (carbon) nanoparticle morphology and their optical properties. [30] They found that open, ramified particles absorb sunlight less efficiently than compact, spherical ones. Other approaches of extracting particle features from images involve additional mathematical transformations, with examples being wavelet [31] and Fourier descriptors,[32, 33] and fractal dimension.[34-36] An interpretation of the latter group of descriptors is less intuitive but may be useful in constructing statistical structure-property models.

While the nanoparticle-QSPR/QSAR modeling methodology is an emerging technology, a major challenge continues to be to select the optimal source of NP structural data. In this contribution, we investigate SEM images as such a source. By applying computer vision algorithms, we quantify the content of SEM images using a set of morphological characteristics. We ask a fundamental question: how to select the optimal images one needs to properly capture both diversity and statistics of the nanoparticles present in a sample set? In the following, we will utilize image analysis algorithms in order to extract descriptors and to perform a series of statistical tests to address these questions. We will illustrate our considerations using images of Tricalcium phosphate ($Ca_3(PO_4)_2$) nanoparticles.

Methods

Images

Tricalcium phosphate (TCP), a member of the calcium orthophosphate family, is a biocompatible and biodegradable compound used both in bulk and nanoparticle form e.g. as a component of composite biomaterials[37]. Herein, we investigate and characterize sub-macro orthophosphate particles obtained from Sigma-Aldrich.

Using Phenom ProX Desktop Scanning Electron Microscope (accelerating voltage: 5,000 – 15,000 V), we obtained 15 grayscale (8-bit) .tiff images of TCP grains. At x400 magnification, the 2,048 by 2,048 pixels (px) micrographs were scaled to 3.061 px/μm (0.3267 μm/px). The number, density and shape of particles varied among the SEM images (Figure 1), and in each case, the objects visible in the images correspond to μm size agglomerates of nanoparticles.

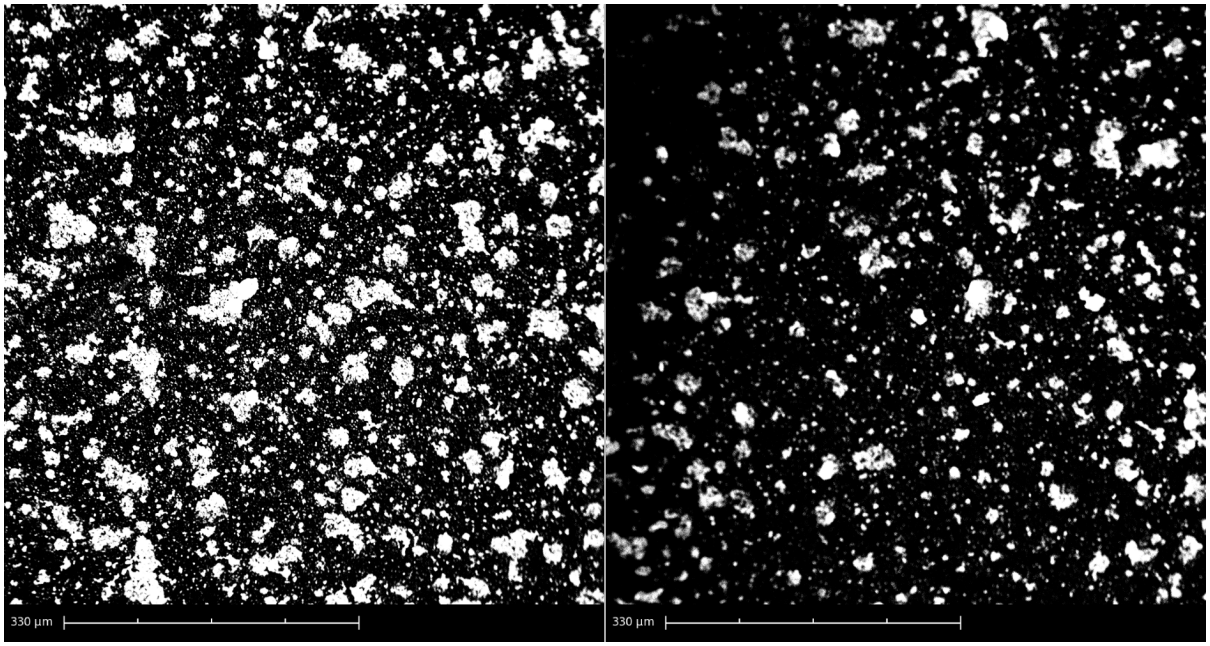

Figure 1. SEM grayscale image samples of tricalcium phosphate (TCP): high variability and pleomorphism.

Image analysis and particle descriptors

Our analysis algorithm was divided into two modules: image processing (preparation) and image analysis (numerical transformation). We use an open source Java-based framework known as ImageJ software to construct our computer vision methodology.[38] ImageJ offers several tools for image enhancement, segmentation and feature extraction, including scripting capability easing the analysis microscopic images.

Image processing and analysis

Prior to any analysis, the SEM images require appropriate preprocessing, such as border improvement and contrast enhancement. In this study, we implemented a workflow with three main processing steps: filtering, thresholding and segmentation. The graphical output (result) at each stage of the procedure is presented in Figure 2, with a brief description given below:

1. FILTERING. Reduce noise, (e.g. dust, dirt, artifacts), through smoothing local variations in the image. Artifact elimination prevents false particle identification during the segmentation stage. *Method:* ***anisotropic diffusion***[39] *- a non-linear filter that homogenize areas with similar intensity while preserving the edges.*

2. THRESHOLDING. Separate the background and the objects based on their pixel intensity (gray-level value, 0-255). Setting an intensity limit aims at eliminating background objects, such as artifacts, specks of adhesive glue, stray fragments of coating film and the like, so that the result contains the foreground objects (aka regions of interest). *Method:* ***intermodes***[40]*, which iteratively attenuates the pixel intensity (gray-level value) variations from the histogram until there are only two maxima, and uses their average as the threshold value for all the pixels in the image .*

3. SEGMENTATION. Identify the foreground objects out of the thresholding step that are most likely to correspond to particles. *Method:* ***discarding*** *objects on edges or with size* ***smaller*** *than* ***10 µm²***

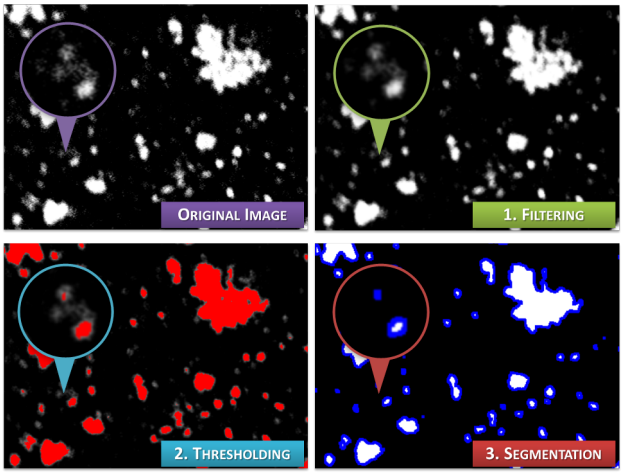

Figure 2. Example of an image processing result throughout the particle identification: colored circles offer a magnified view of the image transformations at each processing step.

Nanoparticle descriptors

After identifying and selecting the particles, our algorithm proceeds in extracting morphological features, which here are scalar numerical descriptors. We obtained ten shape and size descriptors, briefly described in Table 1. The information contained in each image is therefore converted into a list of nanoparticles, and their respective ten descriptors.

Table 1.Calculated morphological particle features: size (orange) and shape (purple) descriptors.

Feature	Description	Unit
Area	area of selection	µm²
Perimeter	length of the outside boundary of the selection	µm
(Ellipse) Major axis	primary axis of the best fitting ellipse; $S_{ellipse} = S_{particle}$	µm

(Ellipse) Minor axis	secondary axis of the best fitting ellipse; $S_{ellipse} = S_{particle}$	µm
Feret's diameter [Max]	maximum distance between the two parallel lines restricting the object perpendicular to a specific direction; maximum caliper	µm
Feret's diameter [Min]	minimum distance between the two parallel lines restricting the object perpendicular to a specific direction; minimum caliper	µm
(Ellipse) Aspect ratio	the ratio of Major Axis to Minor Axis	—
Circularity	comparison of the surface area of a particle to that of a circle with a perimeter of an equal length; $l_{particle} = l_{circle}$	—
Roundness	comparison of the surface area of a particle to that of a circle with a major axis (diameter) of an equal length; $MjrAx_{particle} = MjrAx_{circle}$	—
Solidity	ratio of the particle Area and Convex area; compactness	—

Image representativeness and nanoparticle statistics

Distributions of nanoparticles and their morphology can be significantly different between SEM images. These differences reflect both sample inhomogeneity and deficiencies of sample preparation methods. Therefore, before an image can be used as a source of descriptors for QSAR/QSPR modeling, we must first establish its representativeness, i.e. the level of (morphological) feature diversity expected. A micrograph containing fewer but more diverse particles may be more valuable than one with a large number of particles differing very little from each other. In some cases, it might be necessary to use more than one SEM image in order to incorporate all the possible variations of a descriptor value (particle feature). The challenge is two-fold: (i) to determine a finite number of images, often smaller than the whole set, that will be recorded for later references; (ii) if the visual inspection is mandatory, how to rank the images based on the NP properties?

We developed quantitative protocols for assessing the morphological information about NP carried in SEM images. Employing particle statistics, we designed a series of analytical steps, and implemented algorithms using the R statistical software. The main steps are described as it follows:

STEP 1: Calculating the probability density for each descriptor. The probability density function (PDF), describing the relative likelihood for a variable to take on a specific value, is a tool for assessing the range and frequency of descriptor values. The probability density was calculated using a kernel density estimation (KDE) method implemented in R's *stats* package. These kernels assume a Gaussian function to approximate the data distribution locally, following the equation below:

$$G(x;\ \sigma) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{x^2}{2\sigma^2}}$$

where G is the Gaussian kernel, x is a data point and σ stands for standard deviation of the data.

A set of fifteen SEM images are input to create a reference curve, such that the probability density for each descriptor was estimated on all the particles from the set. In this context, we assume that the available set of images well represent the diversity of the particle sample. In order to estimate PDFs, we assume that the descriptors value range between zero and the respective maximum descriptor value. Within those set boundaries, 1024 evenly spaced points compose the density estimation. The PDF for all the particles in terms of each descriptor serves as reference curve in our convergence analysis.

STEP 2: Determining all the possible image combinations (subsets). In order to investigate potential procedures for optimal image selection, we took all possible image combinations (subsets) under consideration. When listing all possibilities, we started with one-

image subsets, taking into account only one image at a time. Following that, we found all two-image combinations, then all three-image combinations (subsets) and so on – up to 14-image subsets. In total, we found 32,766 possible combinations (subsets) of 15 images.

STEP 3: Calculating subset PDF for each descriptor. Implementing the method described in **STEP 1**, we calculated the PDF of each descriptor for each subset. We used the previously established descriptor bounds (lower and upper range limits) in the subset probability density estimations, thereby ensuring the 1024 sampling point intervals were comparable each time. Additionally, the PDFs here also follow the default implementation, similarly to the smoothing bandwidths for the full set PDF.

STEP 4: Calculating the similarity score. In order to find the difference between the overall (y_i) and subset (f_i) descriptor probability distributions at the *i-th* sample point, we calculated the absolute error (AE):

$$AE = |f_i - y_i| = |e_i|$$

We obtained 10 AE vectors per subset, each for the respective descriptor. Each AE vector contained 1024 values, corresponding to a PDF curve over the sampling intervals. Using the trapezoidal rule[41] we approximated the area under the curve (A_n) in each AE_n vector:

$$A = \int_{x_{min}}^{x_{max}} AE\,(x_i)\,dx \approx (x_{max} - x_{\min})\left[\frac{f(x_{\min}) + f(x_{\max})}{2}\right]$$

As the integral of the probability density function over the entire space is equal to 1, the maximum difference between two PDF curves is equal to 2. Therefore, we normalized all the *A* values and calculated the *similarity score* (*SS*), whereby:

$$SS = 1 - \frac{(A)}{2}$$

The similarity score ranges between 0 and 1. The higher the SS value, the greater the similarity between the overall probability density curve for a specific descriptor and the probability density curve of a given subset.

STEP 5: Performing a two-sample Kolmogorov-Smirnov test (K-S test).[42] This nonparametric test is used for determining whether two sample distributions are statistically significantly different. It is based on the *empirical cumulative distribution function* (*ECDF*)[43]. ECDF assigns a probability of 1/n to each data point, orders the data from smallest to largest in value, and calculates the sum of the assigned probabilities up to and including each data point, hence the term cumulative. The result is a step function, showing increasing probability of obtaining a value smaller than or to equal each given data point in turn:

$$ECDF(x) = \frac{1}{n}\sum_{n=1}^{n} I(x_i \leq x)$$

where x_i is a single data point, I is the indicator function ($I = 1$ for $x_i \leq x$ and $I=0$ for $x_i > x$). The maximum ECDF value in as always 1 the probability of obtaining the complete set of values.

In order to compare the ECDFs of two samples, a maximum distance (D_n) statistic is calculated:

$$D_n = \max_i |ECDF_1(x_i) - ECDF_2(x_i)|$$

The distributions are considered to be different if D_n exceeds a critical value $D_{crit,}$ which can be calculated based on the sample size and significance level α (lowest acceptable probability that the result is an error).

Additionally to D_n, the *p-value* is estimated. It's an alternative way of evaluating the similarity between two ECDFs – assuming that the two samples were taken from the same population, the *p-value* is the probability that the two empirical cumulative distributions would be

as far apart (different) as observed. In other words, the *p-value* is the probability that the maximum distance between the two ECDFs would be greater than or equal to D_n. Small *p-value*, less than or equal to a certain threshold (limit) means that the assumption about the similarity of the two distributions is false. This *p-value* limit is the above-mentioned significance level α.

For ease of comparison, we chose to examine the subset in terms of their *p-values* rather than maximum difference D_n. We chose the standard significance level, α=0.05, meaning we accepted that the results were within a 5% margin of error. When comparing the overall $ECDF_1$ (for the complete 15-maige set) with the $ECDF_2$ of a subset, if $p \leq 0.05$ we concluded that the ECDFs are significantly different, therefore the subset in question is not representative. We applied this *p-value-based* approach to assessing the representatives of all image subsets.

Results

Particle features

Particle number per image ranged with from 414 particles in Image 12 to 951 particles in Image 3, with the average of 635 particles per image; Figure 3 provides a histogram representing distribution of particles in each image.

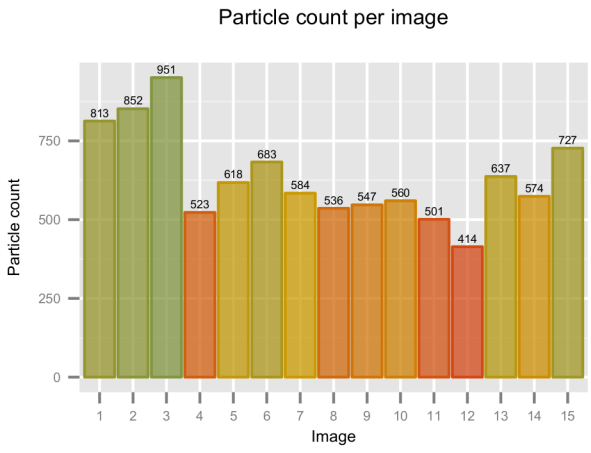

Figure 3. Particle count per image. Bars representing individual images colored from lowest (red) to highest particle count (green).

The majority of the particles are small, under 100 μm^2 (Figure 4: Area) and short, under 10 μm (Figure 4: Ferret's diameter [Max]). The typical shape was close to circular (Figure 4: Circularity) and mostly compact (Figure 4: Solidity). The full set of descriptor value distribution can be found in the Supporting Information.

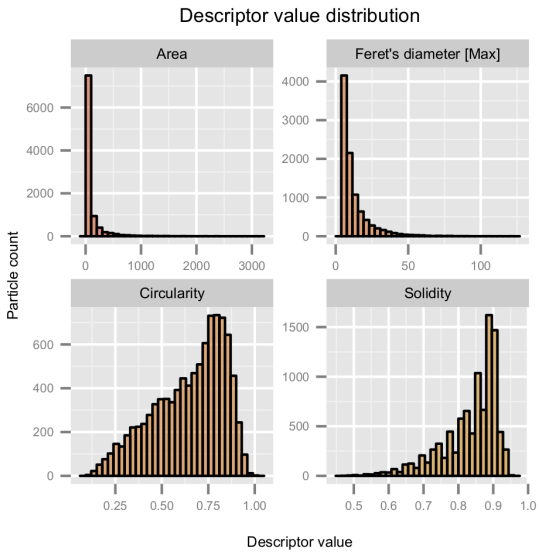

Figure 4. Descriptor value distributions for all images

The section of Image 1 presented in Figure 5 shows how calculated descriptor values correspond to actual particle features. We chose to illustrate the physical representation of two most intuitively interpretable descriptors, one from each of the descriptor groups: the outlines of particles in Figure 5 were colored by Circularity (shape descriptor) and particle fills were colored by Area (size descriptor). As we can see, Circularity values up to 0.5 correspond to elongated particle shapes (Figure 5A, outlined in green and yellow hues) while particles with shapes closer to a perfect circle (Figure 5A, outlined in shades of blue and purple), have higher Circularity

values – between 0.5 and 1. An additional example of particle shapes and corresponding Circularity values (0.25, 0.5, 0.75 and 1) is shown on the left-hand side of Figure 5.

The interpretation of Area descriptor is quite instinctive: the greater the Area value, the larger the actual particle size (Figure 5, smallest particles filled with pink and red, largest particles filled with blue hues).

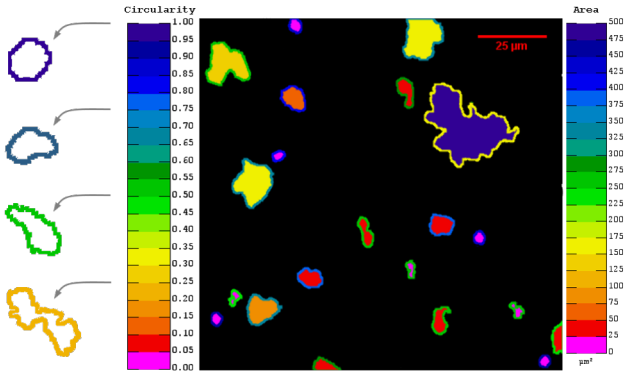

Figure 5. Section of image 1 with particles colored accordingly to their descriptor values: outline representing Circularity, fill representing Area. Polygons on the left-hand side are examples of particle shape (outline) for corresponding Circularity values of: 1.0, 0.75, 0.5 and 0.25.

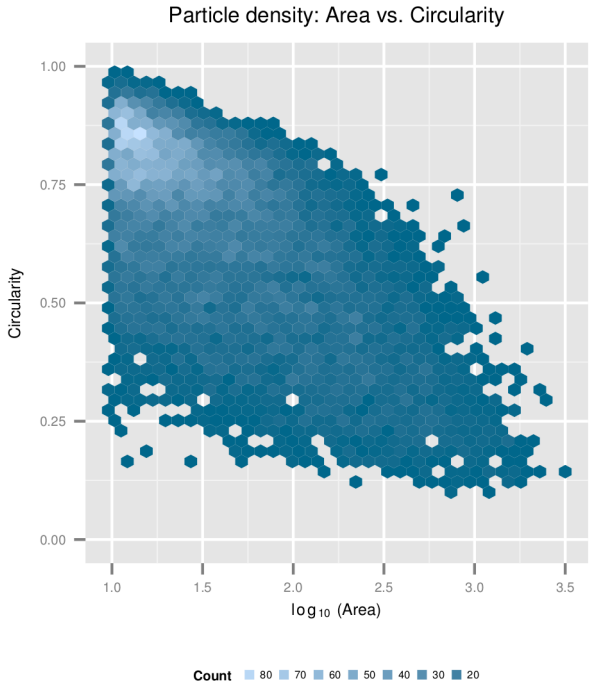

Figure 6. Overall particle count (all images) for Area vs. Circularity. Each hexagon is a two-dimensional bin representing particle count for both descriptors simultaneously.

Further investigation of the relationship between particle's Circularity and their Area revealed an inverse correlation between the two descriptors, illustrated the two-dimensional histogram in Figure 6. Smaller the particles tend have more circular shapes (top-left corner, Figure 6). Conversely, large particles are usually more elongated (bottom-right corner, Figure 6). As descriptor interaction are not the main subject of this work, a full descriptor scatter matrix can be found in the Supporting Information.

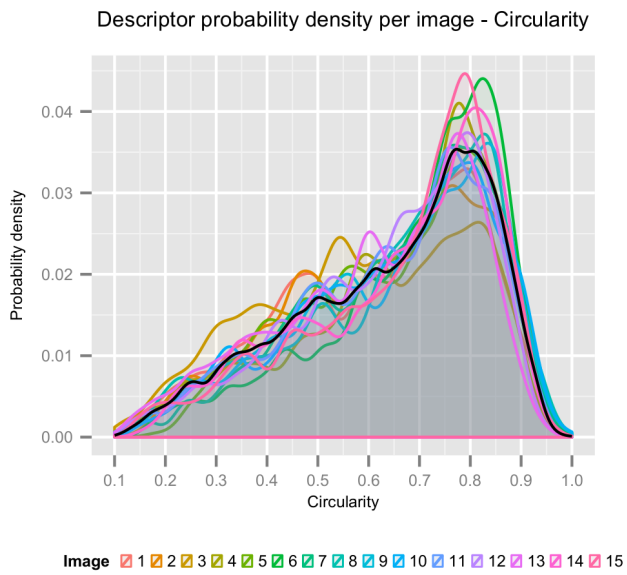

Figure 7. Probability density of descriptor values per image for Circularity. Black line denotes the overall probability density curve (all images at once)

Images vary not only in particle number but also in descriptor value distribution (Figure 7). The overall probability density function (PDF) curve representing all images (black line), and the respective probability density curves for each of the images (various colors) differ significantly. The PDF estimate for Image 7 (turquoise polygon) almost completely overlaps with the overall PDF curve, meaning Image 7 is the most representative in terms of Circularity. Conversely, the PDF curve for Image 3 (yellow polygon) visibly differs from the overall, signifying that image 3 is the least representative of all 15 SEM images. This inter-image variation holds true for other descriptors as well. An overview of the differences in the per-image value distribution for all descriptors can be found in the Supporting Information.

Similarity score

As outlined in the previous section, distributions of particle descriptors differ between images. In this section, we broaden our investigation to include image groups (subsets) and quantify the inter-subset differences with the help of the similarity score.

We observed the same trend for all descriptors: the similarity score values increases with increasing subset size (Figure 8). There is noticeable rise in similarity score: from low values for 1-image subset size group (leftmost, peach-colored bars) to high SS values for 14-image subset size group (rightmost, pink-colored bar). The SS value ranges differ depending on descriptor type: with narrower ranges for size descriptors: 0.8645-0.9968 (Figure 8, panels: 1-6) and wider ranges for shape descriptors: 0.8177-0.9972 (Figure 8, panels 7-10).

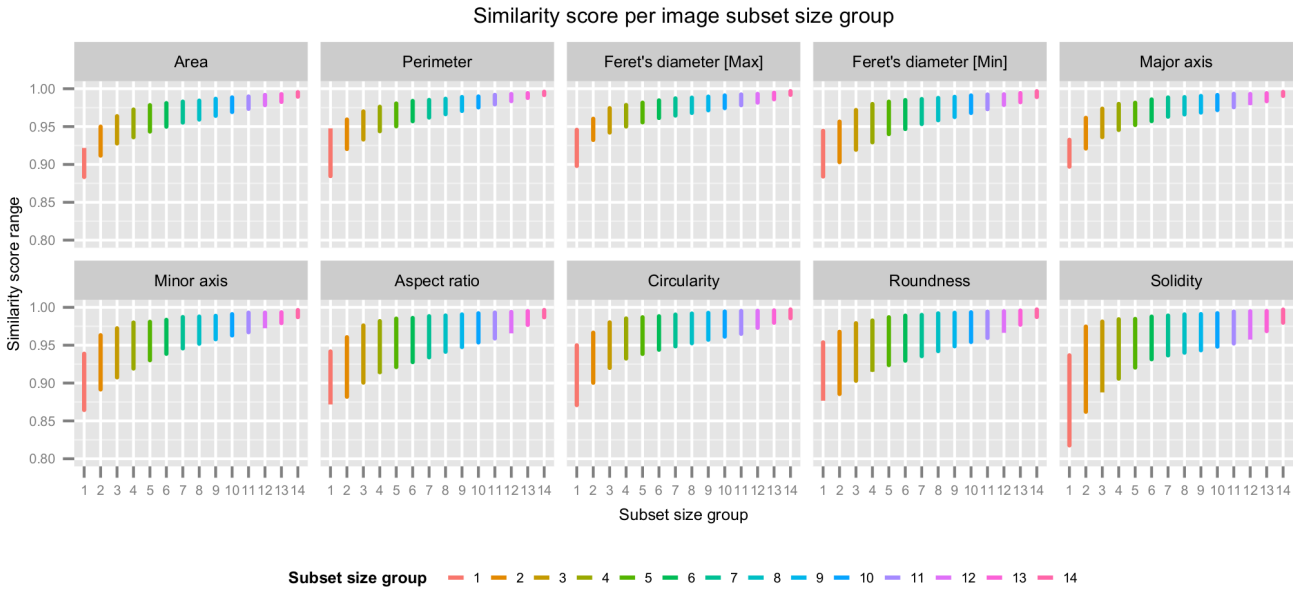

Figure 8. Similarity score ranges per subset size group for all descriptors. Differently colored bars represent SS values ranges for all image subsets of a given size.

We noticed that certain subsets have higher SS values than others of equal size (i.e. containing the same number of images). In the case of Circularity (**Error! Reference source not**

found.), for example, 10-image subsets, represented by blue dots, are spread out vertically - meaning some have lower SS values then others, despite the equal number of images. This difference arises from the fact that these equally sized subsets contain different images, which in turn comprise particles with varying shapes (Circularity). Amongst equally-sized subsets, the ones including certain specific images will be more representative of the whole population (complete image set) than others.

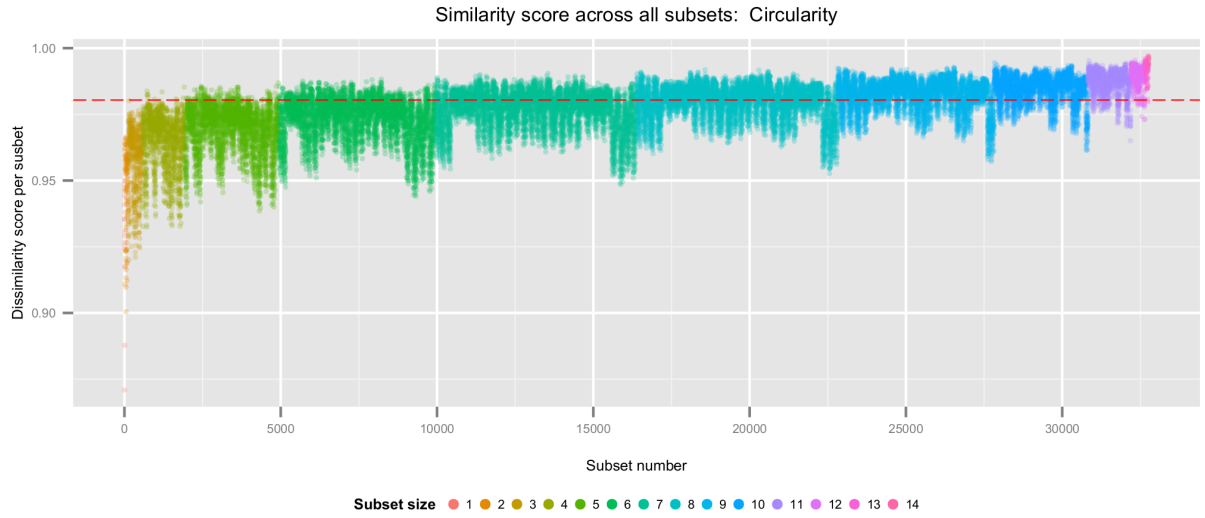

Figure 9. Similarity score of all subsets for Circularity; each dot represents a subset, colored by its size

To further investigate subset quality versus image quantity, we identified the 10-image subset with lowest SS value for Circularity, $SS^{10}_{min} = 0.9228$. This subset included Images 4 thru 10, 12, 14, and 15. Following that, we compared this SS^{10}_{min} with SS values for all smaller subsets (containing 1 through 9 images). We found that subsets as small as three images had SS values higher than 0.9288. One of the 3-image subsets we found, with $SS^{3} = 0.9365$, included only Images 4, 5 and 9. This implies that Images 6, 7, 8, 10, 12, 14 and 15 from the above mentioned

10-image set were in fact redundant – they didn't bring any information that would improve the representatives of the bigger, 10-image, subset. What we can infer from this is that, contrary to our expectations, a larger number of images does not always make a subset more representative. The selection of specific images is equally, if not more, important.

It is worth noting that certain subsets of different sizes have the same SS values - as shown in **Error! Reference source not found.**, there are groups of differently-colors dots with the same SS value. All the dots situated on the red, dashed line represent subsets with SS = 0.9804. This phenomenon occurs in all descriptors. This is further proof that not only the number but also the choice of images has significant influences on a subset's representativeness.

The Kolmogorov-Smirnov test

As mentioned in Section 0, the statistical significance of the differences between the overall and subset's ECDFs can be assessed by comparing the *p-value* against the significance level, α=0.05. The number of subsets with $p \leq \alpha$ (i.e. significantly different from the complete image set) for all descriptors is shown in Table 2, sorted by subset size group. There are fewer significantly different subsets for each of the size descriptors (Table 2, cols. 2 thru 7), e.g. in case of Area, only 22 out of possible 105 2-image subsets are significantly different from the overall set. Whereas in the case of shape descriptors (Table 2, cols. 8 thru 11), there is a significant number of image subset differencing from the complete image set.; e.g. for Solidity, over half of the possible 2-image subsets, that is 60, differ significantly from the complete, 15–image set.

For all but one descriptor, Solidity, is it possible to find a *saturation point*, a specific subset size that, no matter which images are chosen, their combination will never differ significantly from the complete image set. In order to get a representative set of random images for the Area

descriptor for example, we need to use at least 11 out of all 15 images at hand, because at that image number, none of the subsets will differ from the whole set.

Table 2. KS test results: number of significantly different subsets (p-value ≤ 0.05) for all descriptors, sorted by subset size group.

Subset size group	Number of significantly different subsets (p-value ≤ 0.05)										No. of possible subsets
	Area	Perimeter	Feret's diameter Max	Feret's diameter Min	Major Axis	Minor Axis	Aspect ratio	Circularity	Roundness	Solidity	
1	3	1	1	5	1	8	9	5	9	6	15
2	22	5	5	31	8	50	58	39	56	60	105
3	78	25	13	122	28	195	243	186	236	260	455
4	187	49	15	311	48	540	651	559	642	791	1 365
5	299	50	11	537	54	1 028	1 293	1 106	1 270	1 645	3 003
6	338	31	3	705	35	1 424	1 894	1 621	1 876	2 560	5 005
7	254	10	0	638	11	1 484	2 088	1 753	2 057	3 028	6 435
8	119	0	0	416	2	1 127	1 728	1 343	1 706	2 667	6 435
9	30	0	0	190	0	604	1 051	762	1 033	1 692	5 005
10	3	0	0	49	0	224	441	296	429	774	3 003
11	0	0	0	3	0	46	123	75	117	221	1 365
12	0	0	0	0	0	3	18	13	17	44	455
13	0	0	0	0	0	0	2	0	2	5	105
14	0	0	0	0	0	0	0	0	0	1	15
Total	1333	171	48	3007	187	6733	9599	7758	9450	13754	32 766

An example of the differences in ECDF curves between 1-images subsets (i.e. single images) and the overall ECDF curve (black line) for Circularity is presented in Figure 10. We can see that for some of the images (e.g. Images 7 thru 12, teal thru purple hues), the ECDF curves are quite close the overall ECDF (black line) – these are the images with a *p-value* greater than α=0.05 (not significantly different form overall). Whereas the ECDFs of other images, such as Image 3 (yellow line) or Image 6 (green line), are further away from the overall ECDF curve. The *p-value* for those images is smaller than 0.05 and they are considered to be significantly different from the complete image set in terms of Circularity value distribution. The per-image differences in ECDFs for the remaining descriptors have been can be found in the Supporting Information.

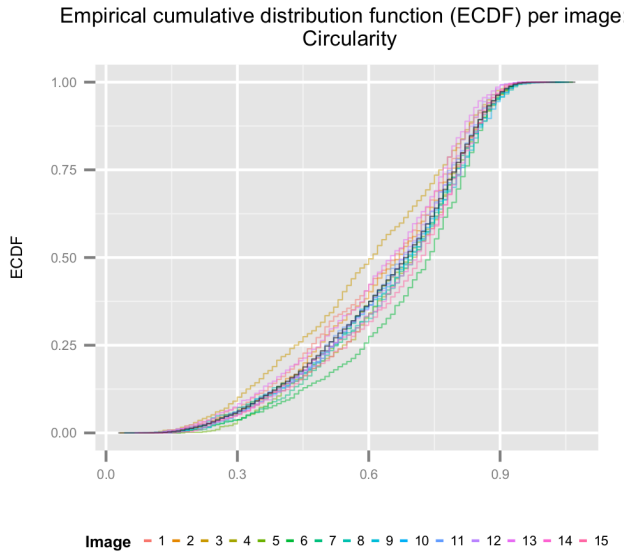

Figure 10. Empirical cumulative distribution function (ECDF) per Image for Circularity. Black line denotes the overall ECDF for the complete image set.

Discussion

Representative image selection

We developed an approach for selecting the most representative images and multi-image sets out of all available images (complete image set) by using the similarity score value a benchmark. For a given image subset size group, the highest SS value will signify the most representative image combination in this group.

In the case of 1-image subsets (i.e. single images), the highest SS score will simply indicate the most representative image (Table 3, highlighted in green) for each descriptor. What is interesting, is that even though some of the images are the most representative for more than one descriptor, e.g. Image 1 (optimal representation of Area and Perimeter) or Image 8 (optimal representation of both Maximum and Minimum Feret's diameter as well as the Minor Axis), there is no single, "globally optimal" image, one that would best represent all of the descriptors at the same time.

Table 3. Similarity score for each of the 15 images from the complete image set and for all descriptors. Highest SS values for each descriptor highlighted in green, lowest values for each descriptor highlighted in red.

Image	Similarity score									
	Area	Perimeter	Feret's diameter [Max]	Feret's diameter [Min]	Major axis	Minor axis	Aspect ratio	Circularity	Roundness	Solidity
1	0.9218	0.9477	0.9344	0.9266	0.9285	0.9145	0.9137	0.9353	0.9248	0.9026
2	0.9097	0.9351	0.9305	0.9259	0.9207	0.9122	0.9097	0.9409	0.9169	0.9162
3	0.9059	0.9351	0.9278	0.9016	0.9147	0.8840	0.8820	0.8709	0.8858	0.8177
4	0.9126	0.9134	0.9160	0.9238	0.9157	0.9144	0.9033	0.9354	0.9071	0.9145
5	0.9108	0.9235	0.9288	0.9298	0.9325	0.9303	0.9352	0.9442	0.9396	0.9229

6	0.9188	0.8999	0.9215	0.9369	0.9277	0.9295	0.9030	0.8878	0.9048	0.8656
7	0.8988	0.9120	0.9316	0.9099	0.9300	0.9127	0.9269	0.9499	0.9361	0.9218
8	0.9157	0.9186	0.9457	0.9444	0.9259	0.9387	0.9244	0.9267	0.9376	0.9268
9	0.9132	0.9295	0.9235	0.9355	0.9313	0.9179	0.9417	0.9290	0.9537	0.9211
10	0.8954	0.9165	0.9309	0.9148	0.9294	0.9097	0.9177	0.9469	0.9252	0.9366
11	0.9187	0.9091	0.9214	0.9071	0.9169	0.9040	0.9259	0.9483	0.9348	0.9327
12	0.8954	0.9118	0.9214	0.8841	0.8972	0.8645	0.8851	0.9305	0.8898	0.9332
13	0.8835	0.8847	0.8981	0.8906	0.9001	0.8861	0.9264	0.9175	0.9303	0.9268
14	0.9117	0.9196	0.9318	0.9225	0.9209	0.9074	0.9130	0.9253	0.9204	0.9345
15	0.8926	0.9184	0.9246	0.9136	0.9208	0.8993	0.8717	0.9239	0.8767	0.9088

Another noteworthy fact is, that despite containing the largest number of particles, Image 3 was not the most representative of any descriptor (Table 3), proving that particle count is not indicative of feature representation and should not be used as a factor during the image selection process.

When selecting representative image (sub)sets, going beyond single images given an opportunity to better represent the entire sample, we can also the use the similarity score value as an indicator.

Table 4. List of Images included in the most representative image combinations of particular size – based on Circularity. Each Image was assigned a unique color for clarity.

Subset size	Max SS	Images included in subset													
1	0.9499	7	-	-	-	-	-	-	-	-	-	-	-	-	-
2	0.9665	2	15	-	-	-	-	-	-	-	-	-	-	-	-
24															

3	0.9801	1	4	9	-	-	-	-	-	-	-	-	-	-
4	0.9853	9	10	13	14	-	-	-	-	-	-	-	-	-
5	0.9869	9	10	11	13	14	-	-	-	-	-	-	-	-
6	0.9882	2	9	10	13	14	15	-	-	-	-	-	-	-
7	0.9905	2	3	6	7	13	14	15	-	-	-	-	-	-
8	0.9918	1	3	5	6	7	8	11	15	-	-	-	-	-
9	0.9928	2	3	6	7	9	10	13	14	15	-	-	-	-
10	0.9945	1	2	3	4	5	6	7	8	12	15	-	-	-
11	0.9953	1	2	3	4	5	6	7	8	11	12	15	-	-
12	0.9955	1	2	3	4	6	7	8	9	10	13	14	15	-
13	0.9960	1	2	3	4	5	6	7	8	9	11	13	14	15
14	0.9971	1	2	3	4	5	6	7	8	9	10	12	13	14

As an illustration, we examine image subsets in terms of how representative they are of particle Circularity. Starting with the smallest subset size group (single images), we observe that Image 7 had the highest similarity score value (Table 4) meaning it is the most representative of the group. When examining the 2-images sets, we find that the combination of Image 2 and Image 15 is the most representative, as this pair had the highest SS value out of all 2-image subsets. When looking for three most representative Images we should combine Images 1, 4, and 9, etc.

Table 5. List of Images included in the most representative image combinations of particular size – based on Area. Each Image was assigned a unique color for clarity.

Subset size	Max SS	Images included in subset
25		

1	0.9218	1	-	-	-	-	-	-	-	-	-	-	-	-	-
2	0.9499	3	13	-	-	-	-	-	-	-	-	-	-	-	-
3	0.9637	3	6	15	-	-	-	-	-	-	-	-	-	-	-
4	0.9725	1	3	7	15	-	-	-	-	-	-	-	-	-	-
5	0.9783	1	3	6	7	15	-	-	-	-	-	-	-	-	-
6	0.9809	1	3	4	7	12	15	-	-	-	-	-	-	-	-
7	0.9830	1	3	5	6	9	12	15	-	-	-	-	-	-	-
8	0.9842	1	3	5	6	7	9	14	15	-	-	-	-	-	-
9	0.9867	1	2	3	5	6	9	11	13	15	-	-	-	-	-
10	0.9885	1	2	3	6	7	8	11	13	14	15	-	-	-	-
11	0.9898	1	2	3	5	6	8	9	10	12	13	15	-	-	-
12	0.9918	1	2	3	5	6	8	9	10	11	12	13	15	-	-
13	0.9929	1	2	3	4	5	6	8	9	10	11	12	13	15	-
14	0.9955	1	2	3	4	5	6	7	8	9	10	12	13	14	15

To assess whether these image combinations would be universally representative, we identified the image subsets with highest SS values in relation to particle Area (Table 5). Interestingly, we found that the best image combinations, those with the highest SS value, we different from the subsets in corresponding size groups for Circularity. Here, Image 1 was most representative of the whole set, whereas for Circularity it had been Image 7. The two most representative image combination was Images 3 and 13, not 2 and 15, as had been the case with Circularity. The three most representative Images in terms of particle Area were 3, 6 and 15 instead of 1, 4, 9 and so on. At each subset size level (form 1-maige subset up to 14-image subset), different sets of Images were representative for either those descriptors.

We further investigated this matter by identifying the most representative (highest SS values) image combinations at each subset size level for eight of the remaining descriptors. We found that the representative image subsets differed from one descriptor to another. While there was some overlap – some subsets were found to have highest SS values for more than descriptor – we did not identify a "universally representative" image combination. For the sake of brevity, we limited the presentation of our findings to the Area and Circularity only.

Minimum representative number of images

Our main goal was to determine the smallest number of images necessary for a representative subset. While the similarity score allows us to identify the most representative images in subset of a given size, it does tell us what the minimum size could be.

As outlined in Section 0, it is possible to find a *saturation point* of a sorts – a subset size that, no matter which images it contains, the subset will always be similar in terms of the specific descriptor distribution to the complete image set. Identifying this *saturation point* for each descriptor will in essence mean finding the smallest representative image set. Each descriptor has a different representative image number, as shown Figure 11. When choosing a set of images representative in terms of Area, we should select at least 11 of them to create a set large enough that even when selecting images as random, we will pick a sufficiently diverse and representative combination. Analogically, in order to obtain a set of random images representative in terms of particle Perimeter, we should use at least 8 of them.

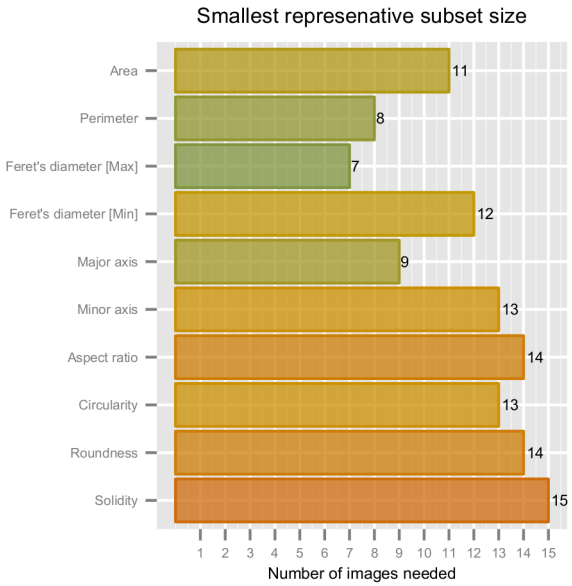

Figure 11. Minimum number of images for a representative image set for each descriptor. Solidity marked I brown as the only representative set is the full set.

This logic applies to all descriptors, with exception of Solidity, for which even among 14-image sets there was still at least one significantly different. This fact signifies that, while our original 15 image set was exhaustive in terms of feature representation for the first nine descriptors (the convergence to the smallest representative image number was always happened under 15 images), the same cannot be said for Solidity. As the minimum representative image set was never found, we cannot be certain that the initial 15 image constituted the *de facto* exhaustive set of images – a larger number might be need to fully represent the diversity in particle Solidity.

Conclusions

We provided a computational framework for extracting morphological features of microparticles from SEM images. Using this framework, we have analyzed 15 SEM images of

tricalcium phosphate TCP material, and investigated their particle distributions in terms of size descriptors: Area, Perimeter, Major and Minor Axes, Maximum and Minimum Feret's diameter, as well as shape descriptors: Aspect ratio, Circularity, Roundness, and Solidity.

We have developed a statistics-based approach for image comparison (similarity score, SS), enabling the selection of most representative images and image sets. We have demonstrated that when choosing a representative set of SEM images, one should make the selection separately for each descriptor. We have proven that information quality (feature diversity) is independent of the number of particles in an image, and the most populated images are not always the optimal choice. We observed an inverse correlation between particle shape and size: smaller are usually circular in shape, while larger particles tend to have more evolved, irregular shapes.

We proposed a method of determining the minimal number of images necessary for a random representative set, with the help of the two-sample Kolmogorov-Smirnov test. Based on the difference in empirical cumulative distribution function (ECDFs) between image subset and the complete image set, we calculated *p-values*. Subsets with *p-values* ≤ 0.05 we classified as significantly different from the complete image set. For every descriptor there is an image number (subset size) large enough, that no matter which random images are combined, the resulting subset will always be representative in terms of that descriptor. Each of the descriptors has a different minimum representative subset size.

To sum up, our algorithms provide a framework for analyzing images of nanoparticle-form materials and selecting minimal sets of images to correctly represent material samples. Our framework is general and could be applied to various materials, using diverse descriptors corresponding to key features, relevant to specific applications. Our methodology could also be

integrated with commercial imaging software, providing immediate, "live" feedback during the image acquisition process.

The next stage of our research will focus on devising a measure of information content for SEM images, based on the distribution of particle features. Following that, we will compare the morphological information obtained by means of computer vision for different biomaterials and group them accordingly to those features, looking for natural clusters. Afterwards, we will investigate relationships between particle morphology and their biological properties.

Supporting Information

Additional figures as well as the set of 15 SEM images are available in the supporting information file.

Acknowledgements:

K.O. and T.P. were supported by the European Commission through the Marie Curie IRSES program, NanoBRIDGES project (FP7-PEOPLE-2011-IRSES, Grant Agreement Number 295128) and by the Foundation for Polish Science (FOCUS Programme). D.U. and M. H. were supported by the Center for Applied Mathematics for Energy Research Applications (CAMERA), funded by the U.S. Department of Energy under Contract No. DE-AC02- 05CH11231. This work for performed at the Berkeley Lab, which operated by the University of California for the U.S. Department of Energy under Contract No. DE-AC02- 05CH11231.

Literature

1. The Chemistry of Nanomaterials. Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, FRG, 2004.

2. Froggett, S. J.; Clancy, S. F.; Boverhof, D. R.; Canady, R. A., A review and perspective of existing research on the release of nanomaterials from solid nanocomposites. Particle and fibre toxicology 2014, 11, 17-17.

3. Zhu, J.; Chen, M.; He, Q.; Shao, L.; Wei, S.; Guo, Z., An overview of the engineered graphene nanostructures and nanocomposites. RSC Advances 2013, 3, 22790-22790.

4. Felder-Flesch, D. (Endo)Fullerenes: From Production to Isolation. In Fullerenes: Principles and Applications (2); The Royal Society of Chemistry: 2012, pp 3-11.

5. Manzano-Ramirez, A.; Moreno-Barcenas, A.; Apatiga-Castro, M.; Mauricio Rivera-Munoz, E.; Nava-Mendoza, R.; Velazquez-Castillo, R., An Overview of Carbon Nanotubes: Synthesis, Purification and Characterization. Current Organic Chemistry 2013, 17, 1858-1866.

6. Pereira, A. E. S.; Grillo, R.; Mello, N. F. S.; Rosa, A. H.; Fraceto, L. F., Application of poly(epsilon-caprolactone) nanoparticles containing atrazine herbicide as an alternative technique to control weeds and reduce damage to the environment. Journal of hazardous materials 2014, 268, 207-15.

7. Singh, P.; Nanda, A., Enhanced sun protection of nano-sized metal oxide particles over conventional metal oxide particles: an in vitro comparative study. International journal of cosmetic science 2014, 36, 273-83.

8. Fang, R. H.; Hu, C.-M. J.; Luk, B. T.; Gao, W.; Copp, J. a.; Tai, Y.; O'Connor, D. E.; Zhang, L., Cancer cell membrane-coated nanoparticles for anticancer vaccination and drug delivery. Nano letters 2014, 14, 2181-8.

9. Global Markets for Nanocomposites, Nanoparticles, Nanoclays, and Nanotubes. BCC Research: Wellesley, 2014.

10. Kumar, A.; Chang, B.; Xagoraraki, I., Human health risk assessment of pharmaceuticals in water: issues and challenges ahead. International journal of environmental research and public health 2010, 7, 3929-53.

11. Hernando, M. D.; Mezcua, M.; Fernández-Alba, a. R.; Barceló, D., Environmental risk assessment of pharmaceutical residues in wastewater effluents, surface waters and sediments. Talanta 2006, 69, 334-42.

12. van de Waterbeemd, H.; Gifford, E., ADMET in silico modelling: towards prediction paradise? Nature reviews. Drug discovery 2003, 2, 192-204.

13. Oomen, A. G.; Bos, P. M. J.; Fernandes, T. F.; Hund-Rinke, K.; Boraschi, D.; Byrne, H. J.; Aschberger, K.; Gottardo, S.; von der Kammer, F.; Kühnel, D.; Hristozov, D.; Marcomini, A.; Migliore, L.; Scott-Fordsmand, J.; Wick, P.; Landsiedel, R., Concern-driven integrated approaches to nanomaterial testing and assessment--report of the NanoSafety Cluster Working Group 10. Nanotoxicology 2014, 8, 334-48.

14. Recent Advances in QSAR Studies. Springer Netherlands: Dordrecht, 2010.

15. Burello, E.; Worth, A. P., QSAR modeling of nanomaterials. Wiley interdisciplinary reviews. Nanomedicine and nanobiotechnology 2011, 3, 298-306.

16. Kar, S.; Gajewicz, A.; Puzyn, T.; Roy, K., Nano-quantitative structure-activity relationship modeling using easily computable and interpretable descriptors for uptake of magnetofluorescent engineered nanoparticles in pancreatic cancer cells. Toxicology in vitro : an international journal published in association with BIBRA 2014, 28, 600-6.

17. Gajewicz, A.; Schaeublin, N.; Rasulev, B.; Hussain, S.; Leszczynska, D.; Puzyn, T.; Leszczynski, J., Towards understanding mechanisms governing cytotoxicity of metal oxides nanoparticles: Hints from nano-QSAR studies. Nanotoxicology 2014, 5390, 1-13.

18. Impoco, G.; Fucà, N.; Pasta, C.; Caccamo, M.; Licitra, G., Quantitative analysis of nanostructures' shape and distribution in micrographs using image analysis. Computers and Electronics in Agriculture 2012, 84, 26-35.

19. Coz, E.; Pérez-Guldris, J.; Calvo, A. I.; Alves, C.; Tarelho, L. A. C.; Ramos, G.; Artiñano, B., A Study on the structural properties of aerosols from biomass combustion for domestic heating. Chemical Engineering Transactions 2014, 37, 811-816.

20. Liati, A.; Dimopoulos Eggenschwiler, P.; Schreiber, D.; Zelenay, V.; Ammann, M., Variations in diesel soot reactivity along the exhaust after-treatment system, based on the morphology and nanostructure of primary soot particles. Combustion and Flame 2013, 160, 671-681.

21. Toth, P.; Farrer, J. K.; Palotas, A. B.; Lighty, J. S.; Eddings, E. G., Automated analysis of heterogeneous carbon nanostructures by high-resolution electron microscopy and on-line image processing. Ultramicroscopy 2013, 129, 53-62.

22. Toth, P.; Palotas, A. B.; Eddings, E. G.; Whitaker, R. T.; Lighty, J. S., A novel framework for the quantitative analysis of high resolution transmission electron micrographs of soot I. Improved measurement of interlayer spacing. Combustion and Flame 2013, 160, 909-919.

23. Toth, P.; Palotas, A. B.; Eddings, E. G.; Whitaker, R. T.; Lighty, J. S., A novel framework for the quantitative analysis of high resolution transmission electron micrographs of soot II. Robust multiscale nanostructure quantification. Combustion and Flame 2013, 160, 920-932.

24. Nine, M.; Choudhury, D.; Hee, A.; Mootanah, R.; Osman, N., Wear Debris Characterization and Corresponding Biological Response: Artificial Hip and Knee Joints. Materials 2014, 7, 980-1016.

25. Moro, T.; Kyomoto, M.; Ishihara, K.; Saiga, K.; Hashimoto, M.; Tanaka, S.; Ito, H.; Tanaka, T.; Oshima, H.; Kawaguchi, H.; Takatori, Y., Grafting of poly(2-methacryloyloxyethyl phosphorylcholine) on polyethylene liner in artificial hip joints reduces production of wear particles. Journal of the mechanical behavior of biomedical materials 2014, 31, 100-6.

26. Wu, J.; Peng, Z., Investigation of the geometries and surface topographies of UHMWPE wear particles. Tribology International 2013, 66, 208-218.

27. Schröder, C.; Reinders, J.; Zietz, C.; Utzschneider, S.; Bader, R.; Kretzer, J. P., Characterization of polyethylene wear particle: The impact of methodology. Acta biomaterialia 2013, 9, 9485-91.

28. Tipper, J. L.; Galvin, A. L.; Williams, S.; McEwen, H. M. J.; Stone, M. H.; Ingham, E.; Fisher, J., Isolation and characterization of UHMWPE wear particles down to ten nanometers in size from in vitro hip and knee joint simulators. Journal of biomedical materials research. Part A 2006, 78, 473-80.

29. Brun, F.; Travan, A.; Accardo, A.; Paoletti, S. Characterization of silver nanoparticles for biomedical applications by means of quantitative analysis of tem micrographs - biomed 2010. 2010; Vol. 46; pp 105-110.

30. Adachi, K.; Chung, S. H.; Buseck, P. R., Shapes of soot aerosol particles and implications for their effects on climate. Journal of Geophysical Research 2010, 115, D15206.

31. Kockentiedt, S.; Toennies, K., Automatic Detection and Recognition of Engineered Nanoparticles in SEM Images. Vision, Modeling & … 2012, 2312-2312.

32. Fernandez Martinez, R.; Okariz, A.; Ibarretxe, J.; Iturrondobeitia, M.; Guraya, T., Use of decision tree models based on evolutionary algorithms for the morphological classification of reinforcing nano-particle aggregates. Computational Materials Science 2014, 92, 102-113.

33. Rice, K. P.; Saunders, A. E.; Stoykovich, M. P., Classifying the Shape of Colloidal Nanocrystals by Complex Fourier Descriptor Analysis. Crystal Growth & Design 2012, 12, 825-831.

34. Florindo, J. B.; Sikora, M. S.; Pereira, E. C.; Bruno, O. M., Characterization of nanostructured material images using fractal descriptors. Physica a-Statistical Mechanics and Its Applications 2013, 392, 1694-1701.

35. Florindo, J. B.; Sikora, M. S.; Pereira, E. C.; Bruno, O. M., Multiscale Fractal Descriptors Applied to Nanoscale Images. Journal of Superconductivity and Novel Magnetism 2012, 26, 2479-2484.

36. Florindo, J. B.; Bruno, O. M., Fractal descriptors based on Fourier spectrum applied to texture analysis. Physica A: Statistical Mechanics and its Applications 2012, 391, 4909-4922.

37. Dorozhkin, S. V., Calcium Orthophosphates in Nature, Biology and Medicine. Materials 2009, 2, 399-498.

38. Schneider, C. a.; Rasband, W. S.; Eliceiri, K. W., NIH Image to ImageJ: 25 years of image analysis. Nature Methods 2012, 9, 671-675.

39. Perona, P.; Malik, J., Scale-space and edge-detection using anisotropic diffusion. IEEE Trans. Pattern Anal. Mach. Intell. 1990, 12, 629-639.

40. Prewitt, J. M. S.; Mendelsohn, M. L., The Analysis Of Cell Images*. Annals of the New York Academy of Sciences 2006, 128, 1035-1053.

41. Jones, O.; Robert, M.; Robinson, A., Introduction to scientific programming and simulation using R. Chapman and Hall/CRC: 2009; Vol. 54.

42. Crawley, M. J., The R Book. 2nd ed.; John Wiley & Sons: Chichester, 2013.

43. Brown, S.; Tauler, R.; Walczak, B., Comprehensive Chemometrics. Elsevier: Oxford, 2009.